GlycopeptideID

User's Guide

Version 0.9 beta

28.02.2014

Hannu Peltoniemi
hannu.peltoniemi@appliednumerics.fi

applied numerics

Contents

Preface 3

Abbreviations 3

1. Introduction 3

2. Workflow 4

    Create Peptide DB & Fetch Proteins 5

    Create Glycan DB & Convert GlycomeDB 5

    Pre Process spectra 5

    Match Peptides 6

    Match Glycans 7

    Estimate False Discovery Rate 7

3. Input formats 8

    MS^2 spectra 8

    Proteins 9

    Glycans 9

4. GlycopeptideID web server 10

5. Result files 11

    Tables (Excel) 11

References 13

Preface

The GlycopeptideID software is developed by Applied Numerics in collaboration with the University of Helsinki. The aim of this document is to give a short overview about the functionality of the GlycopeptideID and to describe the input and output formats. The software is currently under development and it is possible that the actual features may deviate slightly form the ones presented here. For questions about the software you may mail to Hannu Peltoniemi (hannu.peltoniemi@appliednumerics.fi).

Abbreviations

H, Hex:	Hexose
N, HexNAc:	N-Acetylhexosamine
S, NeuAc:	N-Acetylneuraminic acid, sialic acid
F, Fuc:	Fucose, Deoxyhexose
G, NeuGc	N-Glycolylneuraminic acid

1. Introduction

The purpose of the GlycopeptideID tool is to identify intact glycopeptides measured with LC-MS^2 experiments. To accomplish that, it is required that the glycopeptide spectra contain both peptide and glycan fragments. The peptide sequences are identified from the peptide fragments and the glycans from the glycan and glycopeptide fragments. The outcome is a set of best matching peptides and glycans for each MS^2 precursor. The identified glycan can be a glycan composition or a glycan structure (1. order). The tool input is deisotoped MS^2 spectra with identified charge states. The peptides are searched against protein database and glycans against glycan database or *de novo* generated glycan compositions.

The computational methods resample the ones published by Joenväärä *et al.* (2008). The main difference is that in the GlycopeptideID glycan structure database is used and no *de novo* structures are evaluated. Also, the peptide identification is developed further and includes for ex. other protein modifications. Some background to the methods used is also given at Peltoniemi *et al.* (2009, 2013).

2. Workflow

The computational functionality of the GlycopeptideID can be shown as workflow given at the Fig 1. The input is glycopeptide MS^2 spectra which is matched against peptide and glycan databases and/or *de novo* glycans. The program is ran as an automated batch process, i.e. for a given inputs and options the output is generated without interactive steps. For practical reasons the workflow is divided to three separate tasks so that the peptide and glycan database generations are separated from the rest of the workflow.

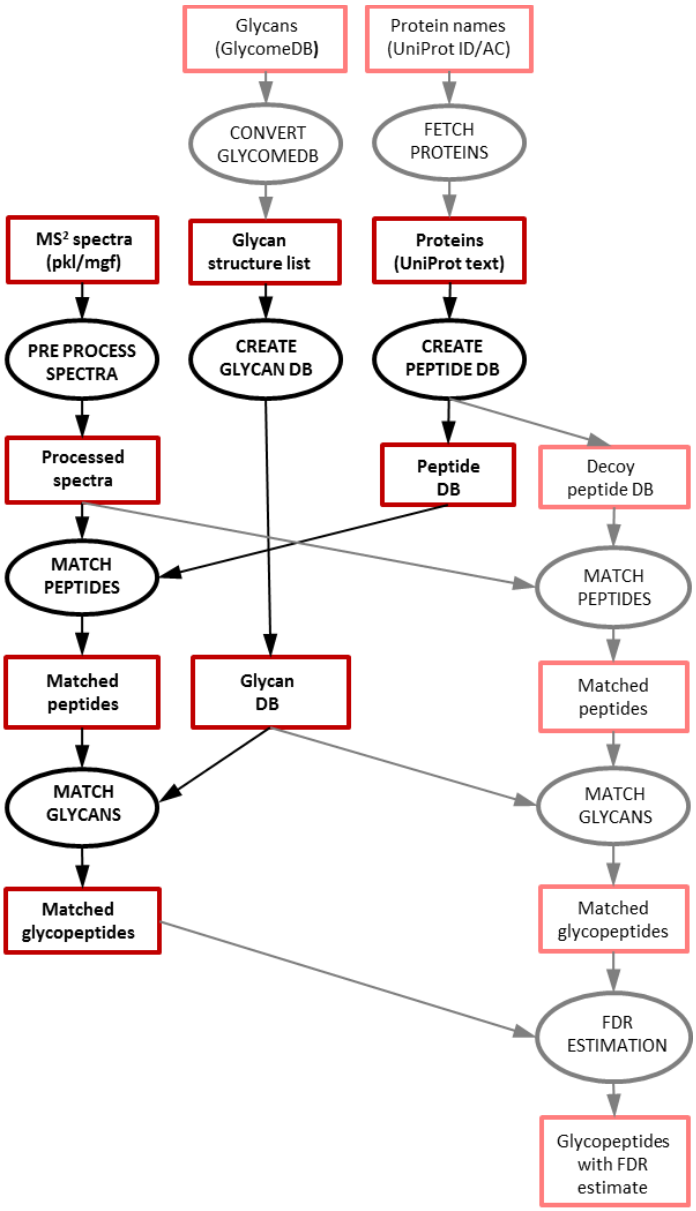

Fig 1. Sketch of the GlycopeptideID workflow. The optional tasks are shown in lighter hue.

Create Peptide DB & Fetch Proteins

Peptide database is created from imported UniProt flat text format protein files (http://www.uniprot.org/). The GlycopeptideID web page contains also a fetch tool suitable for importing a small number of specific proteins directly from the UniProt.

Proteins are digested with trypsin with a given maximum number of missed cleavage sites. Peptides with glycan sites are stored as a peptide database. Glycan sites are defined by a given amino acid rule or by the UniProt database annotation. For false discovery rate (FDR) estimation a decoy database is created by reversing the database peptide sequences.

The peptide database generation have a separate web page and is run separately of the rest of workflow. The reason is that the calculation can be time consuming and typically there is no need to update the peptide database every time the whole workflow is run.

Create Glycan DB & Convert GlycomeDB

Glycan database is created from the GlycomeDB (www.glycome-db.org) database dump xml file, which is pre-loaded to the GlycopeptideID web server. The GlycomeDB file is converted to a structure table with a given taxonomy, glycan core and monosaccharide residues. Glycan fragment database is generated from the structure table with a given maximum number of glycosidic cleavages (ring fragmentation is not included). It is possible to import user defined structure tables and also glycan fragment databases. Importing new or modified databases may be beneficial if there are some special requirements not taken into account with the current service.

As with the peptides the glycan database generation have separate web page and is run separately of the rest of workflow.

Pre Process spectra

The pre-processing contains several optional filtering, combining and tagging tasks.

Filtering:

- Filtering by given precursor RT (retention time), charge or m/z values.
- Filtering by mass tags. The spectra which do not contain any of the typical MS^2 mass tags are filtered. The tags include fragments N, HN, S, SH and SHN. The idea is to get rid of low quality or non glycopeptide spectra which do not contain low mass glycan fragments.
- Filtering by series. The spectra which MS^2 peak differences do not match to any of the monosaccharide (H,N,S,F) masses . As previously the idea is to get rid of low quality or non glycopeptide spectra.

Combining

- MS^2 spectra which match to the same precursor (same charge and RT, m/z differences less than given tolerances) are combined. MS^2 peaks are pasted to a same peak list and intensity values are not summed, so that the resulting spectra contain all the peaks from all the input spectra.

Marking NeuAc tags:

- Spectra with peaks matching to the NeuAc mass tags are marked. The tags include masses of S, SH and SHN. The idea is that if a spectrum matches to a sialylated glycan it should also have fragments matching to the mass tags. Also if the spectrum matches to a non sialylated glycan it should not have matching NeuAc tags.

Match Peptides

Peptide sequences which match best to input spectra are searched. The search has two steps:

1) Optional filter to limit the possible peptides. It is assumed that a MS^2 spectrum contain peaks which can be mapped to a peptide total mass. For example an intact peptide connected to a single HexNAc is typically a major peak in the N-glycopeptide spectrum.

2) Quick sweep to rank peptides with the number of matching fragment ions and calculation of score values for the peptides with highest number of matching fragments. The sweep is done against pre calculated fragment database and is fast enough to be applied even without the filtering step. When calculating peptide scores the intensity is not taken into account as the peptide fragments usually have low intensity compared to glycan and glycopeptide fragments. Details of the score calculation are given at the *Match Glycans* section.

Peptide match can be guided with several options, for example which ion series are searched for, required number of fragment matches etc. It is also possible to limit the output to include only the peptides which can be combined with glycans to match the precursor total mass.

Peptides can be modified. Variable terminus modifications (N- and C-term) can be included with a rather small penalty to the calculation time. Also variable amino acid modifications can be included, but they can be searched only when the peptide total mass filter is applied and they can increase the calculation time considerably. Fixed modifications can be included by modifying the amino acid masses when the peptide database is generated.

Match Glycans

Glycans which match to precursor and potential peptide mass differences are searched and scored against MS^2 spectrum. The search contains steps:

1) Glycans which match to MS^2 precursor and peptide mass difference are searched. Glycans can be structures or compositions. Structures are given as a database and compositions are generated by *de novo* with a given range of monosaccharide units.
2) Glycan fragments (free and attached to peptide) are matched and scored with MS^2 spectra. The theoretical structure fragments for each glycan structure are given in a glycan database. The theoretical composition spectra contain fragments that can be selected from the given composition, i.e. the spectrum is composed by all the fragments that any glycan structure with the given composition could produce.

Scoring the theoretical and measured fragments is done with a statistical score defined by,

(1) $S = -(log_{10}(P_H) + log_{10}(P_I))$,

where P_H is the probability that a random set of fragments would have as many or more shared peaks with the measured spectrum as the ranked glycan and P_I is the probability that by randomly selecting the observed number of shared peaks the same or higher amount of intensity can be covered. The total score for a glycopeptide is given by the sum of glycan and peptide scores.

Note! When scoring a composition it is possible that the matches contain fragments that are difficult to interpret originating from a single credible structure. Especially in case of suspicious compositions and with compositions with nearly equal scores, it is important to validate the matching fragments.

Estimate False Discovery Rate

False discovery rate (FDR) estimation for the peptide sequences of the highest ranking glycopeptides is implemented with the target-decoy approach. The idea is to apply identical workflow (Fig 1.) against target (forward) and decoy (inverted) peptide databases. Random (false) matches are assumed to have an equal probability to match target or decoy, i.e. to have highest scoring peptide in target or decoy databases. The FDR rate is estimated as a (filtered) competition between target and decoy (ref). The precursors are ordered with peptide score and the FDR for matches above a given score value is the percentage of decoy matches relative to target matches. When reporting the matches the decoy hits are assumed to be removed (filtered) from the result.

3. Input formats

The input data contains MS^2 spectra and protein and glycan databases. Examples of all the input data types can be found at the GlycopeptideID Data page.

MS^2 spectra

The supported MS^2 spectra formats are .pkl (Waters) and .mgf (Matrix Science). It is required that the spectra is deconvoluted to a monoisotopic peak lists with resolved charge states. The .pkl format files can be generated with MaxEnt3 algorithm (www.waters.com) and the .mgf files with Mascot Distiller (www.matrixscience.com).

The .**pkl** format file is assumed to contain lines:

[prec m/z] [prec intensity] [prec charge]
[m/z] [intensity]
[m/z] [intensity]
[m/z] [intensity]
..

The first line is precursor ion m/z, intensity and charge and the following lines are fragment m/z and intensity values. The fragments are assumed to be converted to charge 1 peaks (one charge carrier mass included, for ex. one hydrogen). One file can contain several spectra separated be one empty line.

The **.mgf** format is assumed to contain lines:

BEGIN IONS
PEPMASS=[prec m/z] [prec intensity]
CHARGE=[prec charge]+
SCANS=[scan start]-[scan end]
RTINSECONDS=[time start]-[time end]
[m/z] [intensity] [charge] ..
[m/z] [intensity] [charge] ..
...
END IONS

Where *prec m/z*, *prec intensity* and *prec charge* are precursor m/z, intensity and charge values and the *scan start, scan end* and *time start*, *time end* are precursor elution time in scans and seconds. The precursor intensity value is optional and is set to 100 if missing. Also the scan and retention time lines are optional. The precursor data may contain other lines (starting with a character), but are not taken into account.

Lines starting with a number are assumed to contain MS^2 peak data and the first three columns are *m/z*, *intensity* and *charge* values. If the charge is missing the value is set to 1 for each peak. If the charge column is present the table may contain also other columns, but they have no effect on the calculation. Lines starting with BEGIN IONS and END IONS define the beginning and the end of each spectrum.

Proteins

The protein input is the UniProt original flat text format protein files. The GlycopeptideID site contains pre-loaded the entries of the reviewed (Swiss-Prot) human proteins. New protein files can be downloaded from UniProtKB site (www.uniprot.org). When downloading set the file format to text or flat text.

A small set of proteins can also be fetched with GlycopeptideID Protein DB web page. The proteins are identified by UniProt ID names (for ex. TRFE_HUMAN) or by UniProt accessions (for ex. P02787). Each ID (or accession) has to be separated by a new line or comma.

Glycans

The glycan structure tables can be generated from the GlycomeDB (www.glycome-db.org) database dump file or they can be imported by a user. The glycan structure table file is a tab delimited table which contains (at least) three columns: *structure*, *composition* and *source*. The *structure* column contains glycan structures in *modified IUPAC condensed* format using single letter glycan names (H, N, S, F or G) and without linkage (for ex. SHNH(SHNH)HNN). The *composition* contains glycan compositions (for ex. S2H5N4) and *source* is an optional web link to a database where the glycan is fetched from.

The glycan fragment databases can be created from the glycan structure lists. However the fragment database can also be imported by a user. The database is composed by a zipped folder containing a file of glycan structures (glycan_db.txt) and one fragment file for each structure. All files are tab delimited type tables.

The *glycan_db.txt* file contains the three columns present in the structure table and a new column *ms2_theoretical_spectrum_file* which contains the names of the theoretical fragment files for each structure.

The fragment file contains columns: *name*, *structure*, *structure_map*, *composition*, *mass*, *type* and *cuts*. The *name* column contains the ion name and the *structure* is the fragment structure string. The *structure_map* is a string composed by 0's and 1's so that it maps the fragment structure to the whole unfragmented structure string (1: monosaccharide present, 0: absent). The *composition* is fragment composition and the *mass* is the fragment monoisotopic mass. The *type* is the fragment type so that fragments with type 1 are attached to a peptide and the fragments with type 2 are not attached to a peptide. The *cuts* value is the maximum number of glycosidic cuts that are required for a given fragment. The columns *name* and *structure_map* are not required and *composition* and *structure* columns are used only as strings to annotate peaks. The *cuts* column is used to calculate the maximum possible number of dropped H_2O (relative to residual masses) caused by glycosidic cleavages.

4. GlycopeptideID web server

Public version of the GlycopeptideID is found at the www.appliednumerics.com site. The service is available for registered and anonymous (guest) users. For guests there are some limitations for the size and storing time of user data. If there is need for a large scale analysis (several thousand MS^2 spectra), it is possible to set up a dedicated server for the purpose.

The site contains web pages:

Home
- Home page, basic instructions.

Search
- Glycopeptide search options and search start.
- Importing new spectra

Results
- Viewing the calculated results.
- Downloading the results.

Peptide DB
- Generation of a peptide DB.
- Fetching proteins from UniProt.

Glycan DB

- Generation of glycan fragment DB

Data
- Viewing and deleting stored user data

Help
- Documentation

The pages are rather self- explanatory, with some option help and use tips.

5. Result files

All the result files can be downloaded in a single zip file. The file includes some excel templates, an *output* directory containing all the data files and a Results.html file containing links to the main result files.

Tables (Excel)

The main results are mapped to Excel templates glycopeptides.xls and peptides.xls. The templates contain links to *xml* and *pdf* type data which locate at the *output* folder. To open linked xml files with Excel it is required that Excel is set as a default viewer for xml type data (select an xml file with Windows Explorer, select *Open With* → *Choose Default Program* and select Excel).

Peptides (peptides.xsl) contain the matched peptides. The possible columns are:

precursor:
- Running precursor index. The order how the spectra has been in the input file.

mz, intensity, charge:
- Precursor m/z, intensity and charge values.

first_scan, last_scan:
- Precursor retention time start and end scan number.

start_time, end_time:
- Precursor retention time start and end time in minutes

ms2_spectrum_file:
- Link to spectrum data in xml format. If there is a matching molecule the data contains also matching fragments.

combined_precursors:
- Precursor indexes that have been combined to single file. (Optional)

protein:
- Protein name (UniProt ID)

peptide_mass:
- Monoisotopic peptide mass in Da.

start, end:
- Peptide sequence start and end indexes.

N_site:
- N site (N-glycosylation) index.

N_site_DB:
- N site index in UniProt DB.

sequence:
- Peptide sequence.

dmz:
- Mass/charge (m/z) difference between precursor and given molecule.

ion:
- Charge carrier (usually H)

rank_peptide:
- Peptide rank for the given precursor.

score_peptide:
- Peptide score.

hits_peptide:
- Number of matching peptide fragments

I_peptide:
- Relative amount of intensity explained (0:none – 1:all).

b, y
- Matching b and y ion numbers (also ions a,c,x,z if they are matched) .

spectrum:
- Link to pdf picture of MS^2 spectrum with matching fragments.

If modifications are searched the output can contain columns:

modif:
- Modification name (variable aminoacid modification).

modif_site:
- Modification site index.

modif_mass:
- Modification mass value.

modif_AA:
- Modified aminoacid.

modif_number:
- Number of modifications.

N_term_modif:
- N-term modification name.

N_term_modif_mass:
- N-term modification mass.

C_term_modif:
- C-term modification name.

C_term_modif_mass:
- C-term modification mass.

Glycopeptides (glycopeptides.xls) contain the matched glycopeptides. In addition to the peptide columns the possible columns are:

mass_add:
- Additional mass added to composition total mass. (Optional)

composition:
- Glycan composition

structure:
- Glycan structure

ms2_theoretical_spectrum_file:
- Name of the file that contained the theoretical fragments used when matching the structure.

source:
- Glycan source, can be a string or url to a web page.

ppm:
- Glycopeptide and precursor mass difference in ppm.

rank_glycan:
- Glycan rank for given precursor

score_glycan:
- Glycan score.

hits_glycan:
- Number of glycan (glycopeptide) fragment matches.

I_glycan:
- Relative explained intensity by glycan (glycopeptide) fragments (0:none, 1:all).

rank_all:
- Rank defined by total score for given precursor.

score_all:
- Total score, defined as a sum of peptide and glycan scores.

hits_all:
- Total number of hits, sum of peptide and glycan hits.

I_all:
- Total explained intensity, sum of matching peptide and glycan intensities.

decoy_hits:
- Number of decoy hits with a current or higher peptide score value. The value is determined only for best ranking (rank=1) glycopeptides. (Optional)

FDR:
- False discovery rate for peptides of best ranking glycopeptides with a given or higher peptide score. (Optional)

filtered_by_decoy:
- Label (0/1) marking if the precursor has a larger score against decoy (1) or target (0) database. The FDR is calculated with an assumption that the decoy hits are filtered out from the results. (Optional)

References

Joenväärä, S., Ritamo I., Peltoniemi H., Renkonen R., N-Glycoproteomics – an automated workflow approach., Glycobiology. 2008, 18(4):339-349.

Peltoniemi H., Joenväärä S., Renkonen R. 2009. De novo glycan structure search with CID MS/MS spectra of native N-glycopeptides. Glycobiology. 2009, 19:707-714.

Peltoniemi H., Natunen S., Ritamo I., Valmu L., Räbinä J. Novel data analysis tool for semiquantitative LC-MS-MS2 profiling of N-glycans. Glycoconj. J. 2013, 30(2):159-170.